RIP Technique for Frequent Itemset Mining

Shorya Agrawal
GLA University, Mathura, U.P., India

INTRODUCTION

Data mining is a rapidly expanding field being applied in many disciplines, ranging from remote sensing to geographical information systems, computer cartography, environmental assessment and planning. Rule mining is a powerful technique used to discover interesting associations between attributes contained in a database (Han et al., 2006). Association rules can have one or several output attributes. An output attribute from one rule can be used as the input of another rule. Association rules are thus useful, both for obtaining an idea of what concept structures exist in the data (as with unsupervised clustering) and for model creation. In the second instance, the generated rules provide the underlying concepts used in the construction of decision trees and even neural networks, although this is carried out by the automated learning process. Sequential Pattern Mining also comes in Association rule mining For a given transaction database D, an association rule is an expression of the form X $\rightarrow$Y, where X and Y are subsets of attributed set A. The rule X $\rightarrow$Y holds with confidence t, if t% of transactions in D that support X also support Y. The rule X $\rightarrow$Y has support δ in the transaction set D if δ% of transactions in D support X $\cup$ Y. Association rule mining can be divided into two steps. In first step, frequent patterns with respect to support threshold (known as min sup) are mined. In second step, association rules are generated with respect to minimum confidence. Many variants of the ARM based algorithm have been developed.

Proposed technique is termed as Relative Item Path (RIP) based ARM. It creates an innovative graphical structure that is dynamically updated for each transaction in order to determine associated frequent itemset. This technique scores over existing efficient techniques, which had been proposed in recently year. In this proposed technique, each transaction updates the existing graph created by previous transactions, modifying the RIP value associated with the link. The number of scans of database that are required for determining the association of frequent itemset is reduced, saving a great deal of time consumed in database access. The unique feature of the created RIP graph is that it contains nodes equal to total number of items only. This significantly reduces the processing time and memory space required for ARM. The technique works optimally for small and moderate size database. Large databases give rise to enhanced RIP, which are cumbersome in updating. Still, saving in number of access of database and efficient handling of generated RIP graph achieved by the proposed technique make it a strong candidate for determining ARM.

BACKGROUND

The introduction of frequent itemsets (Agrawal et al., 1993), one of the first algorithms proposed for association rules mining was the AIS algorithm. The problem of association rules mining was introduced as well. This algorithm was improved later to obtain the Apriori algorithm (Agrawal et al., 1994).

The Apriori algorithm performs a breadth-first search. Apriori algorithm is a layered search iterative method based on frequency set theory, The core idea is searching for (k+1) item sets through the k item sets, resulting in finding the relationship between database project, in order to form

DOI: 10.4018/978-1-4666-5202-6.ch186

Copyright © 2014, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

the rules. This algorithm includes two steps. The first step is to identify all the frequent item sets, eliminating items having support degree not less than the minimum support degree, which the user specifies; The second step is generation of association rules using frequent itemsets found in step one. In practical applications, Apriori algorithm has some shortcomings. Multiple scanning of database raises time complexity of the algorithm. If the database is very large, it may produce numerous candidates, causing next stage process to become time costlier, which further increases the time complexity. Therefore, in order to improve the efficiency of the Apriori algorithm, a new efficient mining algorithm without candidate set was proposed. Many variants of the Apriori algorithm have been developed, such as AprioriTid, Apriori Hybrid, direct hashing and pruning (DHP), dynamic itemset counting (DIC), Partition algorithm, etc (Agrawal et al., 1993; Srikant et al., 1996). Apriori algorithm based approaches encounter the problem that multiple scans of the database are required in order to determine; which candidates are actually frequent.

Due to the shortcomings of Apriori algorithm another algorithm known as FP tree algorithm was proposed (Han et al., 2000). This approach follows divide and conquer strategy and produces frequent set from FP tree. The highly condensed data structure: FP tree benefits FP-Growth with better performance than the Apriori-like algorithms. It is about an order of magnitude faster than Apriori algorithm. FP Tree data structure is developed for storing patterns from the transaction database. FP-Growth requires two database scans for FP tree construction. During first scan, it finds set of ordered frequent items. A transactional pattern base is constructed during this first scan. During second scan, FP tree is constructed. This Transactional pattern base is substantially smaller in size than the transactional database without loss of any information required for building the FP tree. FP tree is constructed by scanning the transactional pattern base instead of transactional database. This is done by generating first conditional pattern from FP tree. From conditional pattern finally, frequent patterns are extracted. FP-Growth too has some deficiencies. It needs to recursively create huge amounts of conditional pattern bases and corresponding conditional FP tree during mining process. When the dataset is huge, both the memory usage and computational cost are expensive. FP tree is normally smaller in size in comparison with original database. Sometimes even the FP tree itself cannot meet the memory requirement. Because of the huge storage requirement having limited speed of the sequential FP-Growth, parallel algorithm becomes essential for large scale data warehouse mining. Most previous studies (Zaiane et al., 2001; Liu et al., 2007) parallelized the FP-Growth algorithm in a shared memory system.

A related method termed as QFP based ARM, was presented in (Juan et al., 2010). Through scanning the database only once, these algorithm can convert a transaction database into a QFP tree after data pre processing, and then do the association rule mining of the tree. This algorithm performs better than FP-Growth algorithm, and retains the complete information for mining frequent patterns. It does not destroy the long pattern of any transaction and significantly reduce the non-relevant information. Previous approaches in this direction were (Grahne, 2000; Pasquier et al., 2005).

An alternative mining task of mining top-k frequent closed itemsets of length no less than min_l has been taken up in (Wang et al., 2005). Here k is the desired number of frequent closed itemsets to be mined, and min_l is the minimal length of each itemset. An efficient algorithm, called TFP, is developed for mining such itemsets without mins_support. Starting at min_support = 0 and by making use of the length constraint and the properties of top-k frequent closed itemsets, min_support can be raised effectively and FP tree can be pruned dynamically both during and after the construction of the tree using two proposed methods: the closed node count and descendant_ sum. Moreover, mining is further speeded up by employing a top-down and bottom-up combined

8 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/rip-technique-for-frequent-itemset-mining/107394

Related Content

Evaluating the Risks Associated with Supply Chain Agility of an Enterprise
Anirban Ganguly, Debdeep Chatterjeeand Harish V. Rao (2017). *International Journal of Business Analytics (pp. 15-34).*
www.irma-international.org/article/evaluating-the-risks-associated-with-supply-chain-agility-of-an-enterprise/181781

Supply Chain Integration, Collaboration, and Coordination
Genevieve Mushalukand Jing Chen (2014). *Encyclopedia of Business Analytics and Optimization (pp. 2376-2385).*
www.irma-international.org/chapter/supply-chain-integration-collaboration-and-coordination/107421

The Importance of Storytelling in Business Intelligence
Richard T. Herscheland Nicolle Clements (2017). *International Journal of Business Intelligence Research (pp. 26-39).*
www.irma-international.org/article/the-importance-of-storytelling-in-business-intelligence/182763

Causal Feature Selection
Walisson Ferreira Carvalhoand Luis Zarate (2021). *Integration Challenges for Analytics, Business Intelligence, and Data Mining (pp. 145-160).*
www.irma-international.org/chapter/causal-feature-selection/267870

Multi Criteria Decision Model for Risk Assessment of Transmission and Distribution Assets: A Hybrid Approach Using Analytical Hierarchy Process and Weighted Sum Method
Bijoy Chattopadhyayand Angelica Rodriguez (2018). *International Journal of Business Analytics (pp. 33-51).*
www.irma-international.org/article/multi-criteria-decision-model-for-risk-assessment-of-transmission-and-distribution-assets/205642